IEICE TRANS. INF. & SYST., VOL.E106–D, NO.5 MAY 2023

PAPER Special Section on Deep Learning Technologies: Architecture, Optimization, Techniques, and Applications

Multi-Scale Correspondence Learning for Person Image Generation

Shi-Long SHEN[†a)], **Ai-Guo WU**[†b)], ***and*** **Yong XU**[††c)], ***Nonmembers***

SUMMARY A generative model is presented for two types of person image generation in this paper. First, this model is applied to pose-guided person image generation, *i.e.*, converting the pose of a source person image to the target pose while preserving the texture of that source person image. Second, this model is also used for clothing-guided person image generation, *i.e.*, changing the clothing texture of a source person image to the desired clothing texture. The core idea of the proposed model is to establish the multi-scale correspondence, which can effectively address the misalignment introduced by transferring pose, thereby preserving richer information on appearance. Specifically, the proposed model consists of two stages: 1) It first generates the target semantic map imposed on the target pose to provide more accurate guidance during the generation process. 2) After obtaining the multi-scale feature map by the encoder, the multi-scale correspondence is established, which is useful for a fine-grained generation. Experimental results show the proposed method is superior to state-of-the-art methods in pose-guided person image generation and show its effectiveness in clothing-guided person image generation.

key words: *generative models, generative adversarial networks, person image generation*

1. Introduction

Person image generation is regarded as one of the most difficult problems in image analysis and has important applications in movie making, virtual reality, and data enhancement. Pose-guided person image generation, which aims to generate photo-realistic person images based on arbitrary poses, is an important task of this topic. On this challenging task, promising performance has been achieved in some existing methods such as [1]–[9]. For example, a conditional Generative Adversarial Network (GAN) model was recently proposed in [9], which established the correspondence between the input and the exemplar in the feature space to transfer the style from a semantically corresponding region of the exemplar.

Although a meaningful exploration was performed in [9], visual artifacts can be still observed in the generated person images. This may be due to the following reasons: First, only pose keypoints were used as the condition input to synthesize the person images in [9]. However, keypoint-based pose representation is too sparse to accurately represent the correct pose. Second, the problem of misalignment was solved in [9] by establishing the correspondence on a specific scale, and the accurate transfer of information on appearance was achieved. However, a single-scale correspondence may not be able to capture all necessary information for a fine-grained generation. Moreover, there is a limitation in the method of [9], for person image generation, the method can only be applied to pose-guided person image generation. Actually, not only pose but clothing texture can be used to guide the generation process. This task can be called clothing-guided person image generation.

Based on these observations, a generative model is proposed in this paper, which consists of two stages: 1) A pose-guided semantic map generator generates the semantic map guided by the target pose, which allows the model to generate more spatially coherent images. 2) Two pathways are used to get multi-scale features, one for pose encoding and the other for appearance encoding. For the latter, component attributes such as upper cloth, pant, which are separated from the source person image via its semantic map, are used as input to the appearance encoder. After that, the multi-scale correspondence is established in multi-scale correspondence learning (MSC) blocks based on multi-scale features. Finally, the texture renderer equipped with a set of spatially variant de-normalization blocks is used to progressively render the output by using the details from the warp image feature which is obtained based on the multi-scale correspondence.

Experiments are carried out on DeepFashion dataset [10]. Experimental results show that the proposed model can achieve a better result in pose-guided person image generation. Moreover, the proposed method can also implement clothing-guided person image generation, as shown in Fig. 1. The main contributions in this paper are mainly as follows: 1) The multi-scale correspondence between the target pose representation and the source image is established, which can effectively address the misalignment introduced by transferring pose to preserve richer information on appearance. 2) Pose-guided person image generation and clothing-guided person image generation are implemented by using a unified model. 3) The experiment results show that the proposed method is superior to state-of-the-art methods in pose-guided person image generation, and verifies its effectiveness in clothing-guided person image generation.

Manuscript received January 20, 2022.
Manuscript revised March 27, 2022.
Manuscript publicized April 15, 2022.

[†]The authors are with Harbin Institute of Technology (Shenzhen), Shenzhen, China.
[††]The author is with Shenzhen Key Laboratory of Visual Object Detection and Recognition, Shenzhen, China.
a) E-mail: 19s153199@stu.hit.edu.cn
b) E-mail: ag.wu@163.com (Corresponding author)
c) E-mail: yongxu@ymail.com

DOI: 10.1587/transinf.2022DLP0058

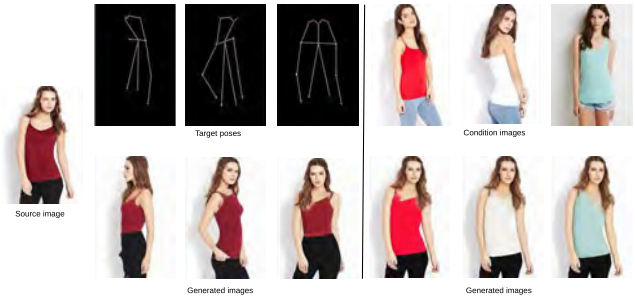

Fig. 1 The proposed method can generate person images in different target poses (left) and transfer upper clothing textures to a person image (right).

2. Related Work

Generative Adversarial Networks (GAN). A GAN [11], consisting of generator and discriminator, has been widely used in the image generation area because of its ability to generate realistic images in adversarial training methods. DCGAN [12] combined convolutional neural network and GAN to generate realistic images through an unsupervised method. In practical applications, conditional images need to be generated. For this end, the so-called CGAN was proposed in [13].

Image to image translation. Image to image translation aims to transfer an image from one domain to another while preserving the original image structure. A typical image translation model is Pix2Pix [14]. Pix2Pix used the conditional Generative Adversarial Networks and took the $\mathcal{L}_1$ loss function as the optimization goal. In addition, in the Pix2Pix model, the concept of PatchGAN was proposed. In this network, a discriminator was utilized to distinguish the authenticity of each image block instead of the authenticity of the entire image. In this design, it is assumeed that the image blocks that are far apart in the image are independent of each other, and the parameters are reduced in the discriminator. However, due to the complexity of human pose transformation and texture, it is not suitable to simply apply the image translation model to the task of person image generation.

Person image generation. One of the difficulties in the task of person image generation lies in how to correctly represent the pose. Keypoint-based pose representation is often used in existing work. Ma *et al*. firstly applied deep learning to the person image generation task in [1]. In the method of [1], a two-stage network was used to generate images with specific pose from coarse to fine. In order to alleviate the problem of deformation in person image generation, a U-net with a deformable skip connection was introduced in [5]. Moreover, PATN in [8] introduced the attention mechanism into the generator to solve the problem of deformation in a progressive generation method. Zhang *et al*. [9] calculated the correspondence between the source image and the target image in the feature space to transfer the appearance information to correspondence position. However, keypoint-based pose representation is too sparse to accurately represent the correct pose, and the generated image suffers from artifacts due to misalignment caused by transferring pose.

The aforementioned methods mainly focus on generating person images based on the input image and target poses. Actually, clothing texture can be also used to guide the process of generation. Inspired by [2], in this paper, the semantic map is used to decouple the source image before obtaining the multi-scale source image features, which allows the proposed method to use a unified model to implement pose-guided person image generation and clothing-guided person image generation.

3. Method Description

In this paper, the task is to achieve person image synthesis. Different from previous pose-guided person image generation methods, clothing-guided person image generation also needs to be considered in the task. To complete this difficult task, the network architecture is divided into two stages: In stage 1, a pose-guided semantic map generator is used to generate the predicted target semantic map. In stage 2, a conditional GAN, consisting of generator and discriminator, is used to generate the output. Specially, during training, given inputs: the target pose P_t, the target semantic map S_t, the source semantic map M_s and the source image I_s, the pose encoder and the appearance encoder are first used to map the pose representation and the decomposed source image to multi-scale feature, respectively. Then, multi-scale correspondence between the target pose representation and the source image can be established in MSC Blocks based on multi-scale pose feature and multi-scale appearance feature, and the warped source image feature can be obtained according to the multi-scale correspondence. Finally, the texture renderer generates the final image based on the warped source image feature and target pose feature. The network training process is shown in Algorithm 1. In the following, firstly, a pose-guided semantic map generator is presented. Then, the generator and discriminator are discussed in detail, respectively. Finally, the objective function used in the present network is described in detail.

3.1 A Pose-Guided Semantic Map Generator

A semantic map can provide more correct pose representation than keypoint-based pose representation during the person image generation process, which has been proven in [15], [16]. In this paper, a semantic map is also adopted as an additional structural constraint. As shown in Fig. 2 (left), the source image I_s, the source semantic map M_s and the target pose P_t are used as the inputs of pose-guided semantic map generator G_{parsing}, which uses a Unet-based network structure, and the predicted target semantic map $\hat{S}_t = G_{\text{parsing}}(I_s, M_s, P_t)$ is generated by minimizing the pixel-wise $\mathcal{L}_1$ loss between S_t and $\hat{S}_t$. Note that if the keypoint-based

Algorithm 1 Network training process

Require:
The source image I_s, the source pose P_s, the source semantic map M_s, the target pose P_t, the target semantic map S_t. Parameters of the pose-guided semantic map generator $\theta_{parsing}$. Parameters of the generator θ_G. Parameters of the discriminator θ_D. Total number of iterative training in stage 1 N_1. Total number of iterative training in stage 2 N_2.

Ensure:
The updated parameters for the pose-guided semantic map generator, the generator and the discriminator.

stage 1:
for epoch=1 to N_1 **do**
  Forward propagation: $\hat{S}_t = G_{parsing}(I_s, M_s, P_s)$;
  Calculate the loss function: $\mathcal{L}_1$;
  Calculate the gradient: $g_{\theta_{parsing}} \leftarrow \nabla_{\theta_{parsing}}[\mathcal{L}_1]$;
  upadte $G_{parsing}$:
    $\theta_{parsing} \leftarrow \theta_{parsing} - \text{Adams}(\theta_{parsing}, g_{\theta_{parsing}})$;
end for
stage 2:
for epoch=1 to N_2 **do**
  Forward propagation for the generator: $I_g = G(P_t, S_t, M_s, I_s)$;
  Calculate the loss function for the generator: $\mathcal{L}_{adv}$, $\mathcal{L}_{fea}$, $\mathcal{L}_{rec}$, $\mathcal{L}_{per}$, $\mathcal{L}_{con}$, $\mathcal{L}_{cor}$, The total loss function of the generator $\mathcal{L}_G$ is the sum of the above loss functions;
  Calculate the gradient: $g_{\theta_G} \leftarrow \nabla_{\theta_G}[\mathcal{L}_G]$;
  update G: $\theta_G \leftarrow \theta_G - \text{Adams}(\theta_G, g_{\theta_G})$;
  Forward propagation for the discriminator: $D(I_t, I_g)$;
  Calculate the loss function for the discriminator: $\mathcal{L}_D$;
  Calculate the gradient: $g_{\theta_D} \leftarrow \nabla_{\theta_D}[\mathcal{L}_D]$;
  update D: $\theta_D :\leftarrow \theta_D - \text{Adams}(\theta_D, g_{\theta_D})$;
end for

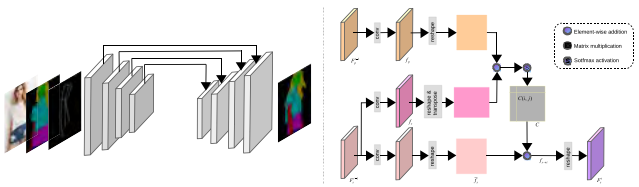

Fig. 2 Left: pose-guided semantic map generator. Right: details of MSC block in the proposed generator.

pose representation is directly used as the input of the generator, the correspondence between the target pose and the source image cannot be established accurately. This problem is addressded in a coarse-to-fine way by predicting the target semantic map firstly. Predicting a target semantic map can not only provide effective structural constraints in the generation process but also be helpful to establish the accurate correspondence between the target pose and the source image.

3.2 Generator

Figure 3 shows the architecture of the generator. During training, the inputs of the generator are the target pose P_t, the target semantic map S_t, the source semantic map M_s and the source image I_s, and the output is the synthesized image $I_g = G(P_t, S_t, M_s, I_s)$ with the texture of I_s and the pose of I_t. At test time, S_t is replaced with $\hat{S}_t$ as input of network since S_t is not available. The target image can be treated as a deformed version of the source image, which means that the pixels on the target image can find the corresponding pixels on the source image. To find the correspondence, the generator encodes the target pose representation and the source image into multi-scale codes by two encoders, called the pose encoder and the appearance encoder. Then, the multi-scale correspondence between the target pose and the source image can be established in MSC blocks. Finally, the texture renderer generates the final result based on the warped source image feature and target pose feature.

Pose encoding and appearance encoding. The pose encoder taking the target pose P_t and the target semantic map S_t as inputs, is used to map the target pose representation into multi-scale codes. Note that the target pose representation is composed of keypoint-based target pose representation and target semantic map.

The appearance encoder takes the source semantic map M_s and the source image I_s as inputs. To implement the task of clothing-guided person image generation, the source semantic map M_s is used to extract correspondence attributes I_s^i of I_s by

$$I_s^i = I_s \odot M_s^i \tag{1}$$

where $\odot$ denotes element-wise product, M_s^i denotes the channel i of M_s. After that, $I_s^i, i = 1 \ldots K$ are concatenated in channel-wise and are used as the input of the appearance encoder. In this way, the proposed model can extract the desired clothing attributes from different source images and combine them to implement clothing-guided person image generation.

MSC blocks. **M**ulti-**s**cale **c**orrespondence learning blocks (MSC blocks), consisting of multi-MSC block as shown in Fig. 2 (right), are used to establish the multi-scale correspondence between the target pose represtation and the source image. Specifically, $F_p^{t-1} \in \mathbb{R}^{c \times h \times w}$ represents the pose feature and $F_s^{t-1} \in \mathbb{R}^{c \times h \times w}$ represents the appearance feature, where h, w are feature spatial size and c is the channel wise demension. F_p^{t-1} and F_s^{t-1} are feed into a convolution layer to generate $f_p \in \mathbb{R}^{c \times h \times w}$ and $f_s \in \mathbb{R}^{c \times h \times w}$, respectively. Then, f_p is reshaped to $\mathbb{R}^{(hw) \times c}$ and f_s is reshaped to $\mathbb{R}^{c \times (hw)}$. After that, the correspondence between the target pose represtation and the source image can be established by

$$\mathbf{C}(i,j) = \frac{(f_p(i) - \mu(f_p)) \cdot (f_s(j) - \mu(f_s))}{\|f_p(i) - \mu(f_p)\|_2 \cdot \|f_s(j) - \mu(f_s)\|_2} \tag{2}$$

where $\mu(f_p)$ and $\mu(f_s)$ represent mean value. $\mathbf{C} \in \mathbb{R}^{hw \times hw}$ is called correspondence matrix, whose element $\mathbf{C}(i,j)$ measures the similarity of f_p at point i and f_s at point j. Then, the correspondence matrix can be used to warp the source image feature by

$$f_{s \to t}(i) = \sum_j \underset{j}{\text{softmax}}(\mathbf{C}(i,j)) \cdot \bar{f}_s(j) \tag{3}$$

where $\bar{f}_s \in \mathbb{R}^{(hw) \times c}$ is obtained by applying convolution and reshape operations, and $f_{s \to t} \in \mathbb{R}^{(hw) \times c}$. Finally, $f_{s \to t}$ is reshaped to $F_s^t \in \mathbb{R}^{c \times h \times w}$. F_s^t is the feature of source image

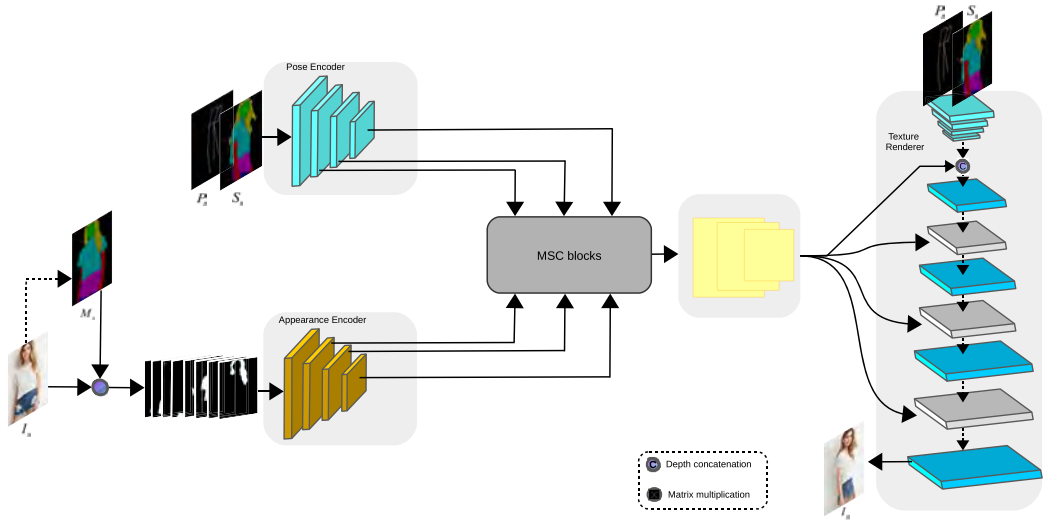

Fig. 3 An overview of the network architecture of the generator.

after deformation, which contains target pose information. Note that different MSC block is used for features of different scales. Here, the method of [9] is used for reference to establish the correspondence between the target pose representation and the source image. Unlike [9], which establishes the single-scale correspondence, the multi-scale correspondence is establish in the proposed model. It is shown in the experiment that the single-scale correspondence is not enough to capture all necessary information and establish accurate correspondence for our complex task, thus the multi-scale correspondence is explored for a fine-grained generation and the experimental results illustrate its effectiveness.

Texture rendering. After obtaining the warped source image feature, the feature needs to be used to generate the final output. Figure 3 shows the architecture of the texture renderer. To better preserve the information of texture feature, the SPADE network structure [17] is borrowed. The texture features of different sizes are used as the input of the SPADE module. But some changes have been made to the SPADE network structure: 1) As opposed to [17], the BN layer in the SPADE module is replaced with the IN layer, because in the image generation task, different images have different styles, therefore, the IN layer that performs feature statistics on a single channel of a single instance is more suitable. 2) The input of texture renderer is replaced from the constant code to the target pose feature. Compared with the constant code, the target pose feature has richer information, which can speed up the convergence of the model. Moreover, the target pose feature can provide position constraints for texture feature, leading to a more realistic generated image.

3.3 Discriminator

Inspired by [18], GAN loss and feature loss are combined in the discriminator to achieve a stable training effect. The inputs of the discriminator are the real image and the generated image. The feature loss function is calculated at the output of each layer of the discriminator and the GAN loss function is calculated in the last layer.

3.4 Objective Functions

Due to the complexity of the task, a joint loss function is proposed to train the proposed network. It repreesnts a combination of the adversarial loss, feature loss, reconstruction loss, perceptual loss, contextual loss and correspondence loss, written as follows:

$$\begin{aligned}\mathcal{L}_{\text{total}} &= \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{fea}}\mathcal{L}_{\text{fea}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} \\ &+\lambda_{\text{per}}\mathcal{L}_{\text{per}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{cor}}\mathcal{L}_{\text{cor}}\end{aligned} \tag{4}$$

where λ_{adv}, λ_{fea}, λ_{rec}, λ_{per}, λ_{con}, λ_{cor} denote the weights of corresponding losses. The goal of adversarial loss is to make the distribution of the generated image as close as possible to the distribution of the real image, which is defined as

$$\mathcal{L}_{\text{adv}} = \mathbb{E}[\log(1-D(G(I_{\text{s}}, S_{\text{s}}, S_{\text{t}}, P_{\text{t}}))] + \mathbb{E}[\log D(I_{\text{t}})] \tag{5}$$

The feature loss can be written as

$$\mathcal{L}_{\text{fea}} = \sum_{i=0}^{n} \alpha_i \| D_i(I_{\text{g}}) - D_i(I_{\text{t}}) \|_1 \tag{6}$$

where D_i denotes the i−th, $i = 0, 1, 2$ layer feature from discriminator, α_i denotes the weight of the feature loss of each layer. The reconstruction loss is used to penalize the difference between the generated image and the real image at the pixel level:

$$\mathcal{L}_{\text{rec}} = \| I_{\text{g}} - I_{\text{t}} \|_1 \tag{7}$$

In addition, the perceptual loss

$$\mathcal{L}_{\text{per}} = \| \phi_l(I_{\text{g}}) - \phi_l(I_{\text{t}}) \|_1 \tag{8}$$

Table 1 The detailed architecture of our approach. k7s1p3 indicates the convolutional layer with kernel size 7, stride 1 and padding 3.

Module	Network layer	Output shape(H×W×C)
pose encoder / appereance encoder	Conv2d / k7s1p3 + Resblock / k3s1	256× 256 × 3
	Conv2d / k4s2p1 + Resblock / k3s1	128× 128 × 16
	Conv2d / k4s2p1 + Resblock / k3s1	64× 64 × 32
	Conv2d / k4s2p1 + Resblock / k3s1	32× 32 × 64
	Conv2d / k4s2p1 + Resblock / k3s1	16× 16 × 128
MSC blocks	MSC block×3	64× 64 × 32
		32× 32 × 64
		16× 16 × 128
Texture rendering	SPADE ResBlk	16× 16 × 128
	Bilinear Interpolation	32 × 32 × 128
	SPADE ResBlk	32× 32 × 64
	Bilinear Interpolation	64 × 64 × 64
	SPADE ResBlk	64 × 64 × 16
	Bilinear Interpolation	128 × 128 × 16
	SPADE ResBlk	128 × 128 × 3
	Bilinear Interpolation	256 × 256 × 3
	Conv2d / k3s1p1	256 × 256 × 3
	tanh	256 × 256 × 3
Discriminator	Conv2d / k7s1p3 + Resblock / k3s1	256 × 256 × 3
	Conv2d / k4s2p1 + Resblock / k3s1	128 × 128 × 16
	Conv2d / k4s2p1 + Resblock / k3s1	64× 64 × 32

where ϕ_l denotes the output of $l-$th layer from the pretarined VGG-19 model, is also used to match the deep features of the image and is effective in image generation tasks. We also adopt contextual loss proposed in [19], which is designed for image generation that naturally handles tasks with non-aligned training data and is very suitable for the task in the current paper. The contexttual loss is used:

$$\mathcal{L}_{\text{con}} = -\log(\text{CX}(\phi_l(I_t), \phi_l(I_g)) \tag{9}$$

where CX denotes the contexttual similarity between the feature of synthesized image and the feature of target image. The detailed definition can be found in [19]. Finally, in order to improve the similarity between F_s^t and the target image in the feature space, the following cost function is used:

$$\mathcal{L}_{\text{cor}} = \|F_{\text{s}}^{\text{t}} - \phi_l(I_{\text{t}})\|_1 \tag{10}$$

4. Experiments

4.1 Implementation Details

Datasets. Experiments are carried out on the In-shop Clothes Retrieval Benchmark of the Deepfashion dataset [10], which contains 52,712 images of people with varying poses and appearances. Adopting the data division configuration in [2], 10,1966 image pairs are used as training dataset and 8750 image pairs are used as test dataset, to ensure that there is no overlap between the two sets.

Metrics. Inception Score (IS) [20], Structural Similarity (SSIM) [21], Frechet Inception Distance (FID) [22] and Learned Perceptual Image Patch Similarity (LPIPS) [23] are used to quantitatively evaluate the quality of the generated image, which are commonly used as evaluation metrics in image generation task. For IS and SSIM, a higher score is better. For FID and LPIPS, a lower score is better.

Network Implementation and Training Details. The proposed network is implemented based on PyTorch, using four 2080Ti GPUs. The pose encoder and the appearance encoder both contain 4 down-sampling layers, and the texture renderer is composed of 4 SPADE ResBlks. The discriminator is composed of three down-sampling layers. The detailed parameter setting is shown in Table 1. In each layer of the network, spectrum normalization [24] is used to stabilize network training. The Adams optimizer [25] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ is used. Inspired by TTUR [22], the initial learning rates of the generator and discriminator are set to 0.0002 and 0.0003, respectively. The batch size is set to 4, the training process lasts for 30 epochs, and the learning rate is linearly decayed to 0 after 15 epochs. The weights for the loss terms in (4) are set to λ_{adv}=1, λ_{fea}=1, λ_{rec}=0.01, λ_{per}=0.5, λ_{con}=0.4, λ_{cor}=1.

4.2 Pose-Guided Person Image Generation Results

The results of pose-guided image synthesis are shown in Fig. 4. Given a source image and any target pose, the proposed method can convert the pose to the target pose while maintaining the texture of the source image.

4.2.1 Comparison with SOTA

For pose-guided person image generation, the proposed method is compared with five state-of-the-art methods, *i.e.*, PG2 [1], DefGAN [5], PATN [8] and ADGAN [2], CoCos-

Fig. 5 Qualitative comparison with state-of-the-art methods.

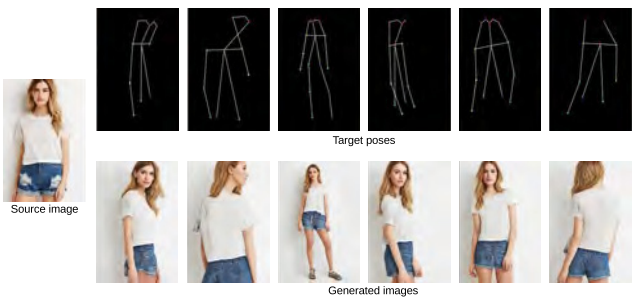

Fig. 4 Results of synthesizing person images in arbitrary poses from the Deepfashion dataset

Net [9]. The results of the comparison methods are obtained through the open-source code and pre-trained models released by the authors of that papers.

Qualitative comparison. Among the comparison methods, Def-GAN, PATN, and CoCosNet take the deformation between the source image and the target image into consideration but PG2 and ADGAN don't. The qualitative comparison result is shown in Fig. 5. It can be learnd from the results that compared with PG2 and ADGAN, the proposed method has the correct pose and more detailed texture results due to the consideration of the deformation between the source image and the target image. On the other hand, compared with other methods, such as DefGAN, PATN, and CoCosNet, the proposed method can also generate more natural and realistic results due to the multi-scale correspondence, especially on face identity and hair texture.

Quantitative comparison. To verify the effectiveness of the proposed method more objectively, quantitative experiments are conducted to compare the proposed method with state-of-the-art methods. The results of the quantitative comparison are shown in Table 2. It can be known from the table that the proposed method is better than state-of-the-art methods in all four evaluation metrics, increasing the best IS score from 0.773 to 0.813, improving the best SSIM score from 3.439 to 3.542, reducing the best FID score from 13.009 to 11.307 and dropping the best LPIPS score from 0.177 to 0.127. On the other hand, the proposed method has fewer parameters than PG2 [1], Def-GAN [5] and CoCosNet [9]. The results of qualitative experiment further verify the effectiveness of the proposed method.

4.2.2 Ablation Study

To verify the influence of the important part of the proposed method on the final result, the ablation study is conducted. The ablation study is divided into the following parts: **w/o parsing:** The semantic map is not used as target pose representation, but only keypoint-based pose representation. **w/o pose feature:** The input of the texture renderer is replaced from the target pose feature to the constant code. **w/o GAN loss:** The proposed model is trained without using GAN loss. **w/o MSC blocks:** The single-scale correspondence

Table 2 Quantitative comparison with state-of-the-art methods and ablation study. "Deform" indicates whether the method models deformation and "M" indicates millions.

Model	Deform	SSIM ↑	IS ↑	FID ↓	LPIPS ↓	Parameters
PG2 [1] (NIPS2017)	✗	0.773	3.202	47.713	0.245	437.09M
Def-GAN [5] (CVPR2018)	✓	0.756	3.439	26.430	0.209	82.08M
PATN [8] (CVPR2019)	✓	0.771	3.203	19.822	0.196	**41.36M**
ADGAN [2] (CVPR2020)	✗	0.770	3.392	13.009	0.177	48.79M
CoCosNet [9] (CVPR2020)	✓	0.759	3.280	15.022	0.194	145.50M
w/o parsing	✓	0.727	3.461	45.404	0.242	-
w/o pose feature	✓	0.806	3.410	13.487	0.134	-
w/o GAN loss	✓	0.808	3.448	13.439	0.132	-
w/o MSC Blocks	✓	0.811	3.507	11.874	0.136	-
Ours(full)	✓	**0.813**	**3.542**	**11.307**	**0.127**	59.60M
Real Data	-	1.000	4.053	0.000	0.000	-

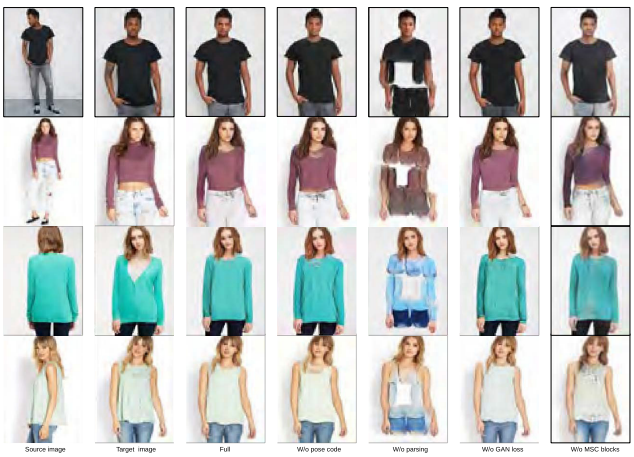

Fig. 6 Qualitative results of the ablation study.

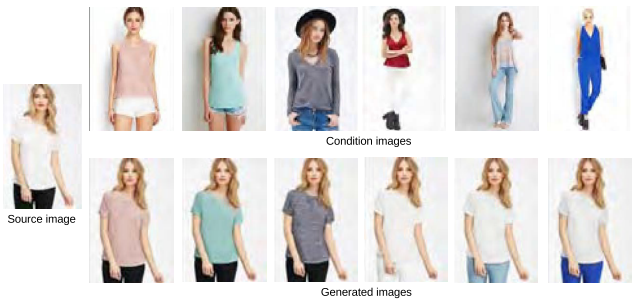

Fig. 7 Results of synthesizing person images with controllable component attributes.

is established between the target pose representation and the source image. The results of the ablation study are shown in Table 2 and Fig. 6. It can be known from the results that the target semantic map can provide effective structural constraints during the image generation process. Moreover, the target pose feature as input can effectively guide the rendering of texture features, and under the constraint of GAN loss, the network can generate a more realistic image. From the comparison result, it can be seen that the multi-scale correspondence learning can improve the quality of the generated image.

4.3 Clothing-Guided Person Image Generation Results

As described in Sect. 3.2, the proposed model can extract the desired clothing attributes from different source images and combine them to implement clothing-guided person image generation. As shown in Fig. 7, for the given source image, the first row represents the conditional image with the desired clothing attributes, and the second row represents the generated image. In the first three columns, the proposed model can change the upper clothes of the source image according to the desired clothing attributes. In the last three columns, the proposed model can change the pants of the source image according to the desired clothing attributes.

5. Conclusion

In this paper, a generative model is proposed for person image generation, and is applied to pose-guided person image generation and clothing-guided person image generation. At the core of our model is the multi-scale correspondence learning between the target pose representation and the source image, which effectively addresses the misalignment introduced by transferring pose to preserve richer information on appearance. Experimental results illustrate that the proposed method is better than state-of-the-art methods in pose-guided person image generation and show its effectiveness in clothing-guided person image generation. In the future work, the approach presented in the current paper will be extended to person video generation. The key difficulty of this issue is how to ensure the timing consistency between frames. Moreover, we will try to use other complex data set and improve the generalization ability of the model. Finally, person image generation in complex backgrounds is also an issue worthy of investigation.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Project No.

2018AAA0100102 and by Innovation and Entrepreneurship Team Project of Chaozhou with Contract No. 220217157150517.

References

[1] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," Proc. Conference and Workshop on Advances in Neural Information Processing Systems, Long Beach, America, pp.406–416, Dec. 2017.

[2] Y. Men, Y. Mao, Y. Jiang, W.-Y. Ma, and Z. Lian, "Controllable person image synthesis with attribute-decomposed gan," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, pp.5084–5093, June 2020.

[3] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, America, pp.99–108, June 2018.

[4] P. Esser and E. Sutter, "A variational u-net for conditional appearance and shape generation," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, America, pp.8857–8866, June 2018.

[5] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable gans for pose-based human image generation," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, America, pp.3408–3416, June 2018.

[6] Y. Ren, X. Yu, J. Chen, T.H. Li, and G. Li, "Deep image spatial transformation for person image generation," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, pp.7690–7699, June 2020.

[7] Y. Li, C. Huang, and C.C. Loy, "Dense intrinsic appearance flow for human pose transfer," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, America, pp.3693–3702, June 2019.

[8] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, America, pp.2347–2356, June 2019.

[9] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, "Cross-domain correspondence learning for exemplar-based image translation," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, pp.5143–5153, June 2020.

[10] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, America, pp.1096–1104, June 2016.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Proc. Conference and Workshop on Advances in Neural Information Processing Systems, Montréal, Canada, pp.2672–2680, Dec. 2014.

[12] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," Proc. 4th International Conference on Learning Representations, Puerto Rico, America, pp.1–16, May 2016.

[13] M. Mirza and S. Osindero, "Conditional generative adversarial nets," CoRR, vol.abs/1411.1784, 2014.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A.A. Efros, "Image-to-image translation with conditional adversarial networks," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Hawaii, America, pp.1125–1134, July 2017.

[15] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated warping-gan for pose-guided person image synthesis," Proc. Conference and Workshop on Advances in Neural Information Processing Systems, Montréal, Canada, pp.474–484, Dec. 2018.

[16] X. Han, X. Hu, W. Huang, and M.R. Scott, "Clothflow: A flow-based model for clothed person generation," Proc. IEEE/CVF International Conference on Computer Vision, Seoul, Korea, pp.10471–10480, Oct. 2019.

[17] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, America, pp.2337–2346, June 2019.

[18] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Hawaii, America, pp.8798–8807, June 2018.

[19] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," Proc. European Conference on Computer Vision, Munich, Germany, pp.768–783, Sept. 2018.

[20] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," Proc. Conference and Workshop on Advances in Neural Information Processing Systems, Barcelona, Spain, pp.2234–2242, Dec. 2016.

[21] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. Image Process., vol.13, no.4, pp.600–612, 2004.

[22] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," Proc. Conference and Workshop on Advances in Neural Information Processing Systems, Long Beach, America, pp.6626–6637, Dec. 2017.

[23] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, America, pp.586–595, June 2018.

[24] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," Proc. 6th International Conference on Learning Representations, Vancouver, Canada, pp.1–26, April 2018.

[25] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Proc. 3rd International Conference on Learning Representations, California, America, pp.1–15, May 2015.

Shi-Long Shen received eceived his B. Eng. degree in Automation in July 2019 from Harbin Institute of Technology, He is currently pursuing the M. Eng. degree with the Harbin Institute of Technology (Shenzhen). His research interest includes deep learning, Human Image Processing.

Ai-Guo Wu received his B. Eng. degree in Automation in July 2002, M. Eng. degree in Navigation, Guidance and Control in July 2004, and Ph.D. degree in Control Science and Engineering in November 2008 all from Harbin Institute of Technology. In October 2008, he joined Harbin Institute of Technology Shenzhen Graduate School, where he is now a professor. Prof. Wu visited City University of Hong Kong from March 2009 to March 2011 as a Research Fellow. His research interests include descriptor systems, conjugate product of polynomials, switched systems. Prof. Wu is a Reviewer for American Mathematical Review. He was an Outstanding Reviewer for IEEE Transactions on Automatic Control. He received the National Natural Science Award (Second Prize) in 2015 from P. R. China, and the National Excellent Doctoral Dissertation Award in 2011 from the Academic Degrees Committee of the State Council and the Ministry of Education of P. R. China. He was supported by the Program for New Century Excellent Talents in University in 2011.

Yong Xu received the B.S. and M.S. degrees in 1994 and 1997, respectively, andthe Ph.D. degree in pattern recognition and intelligence system from NUST, China, in 2005. He is currently working with the Harbin Institute of Technology, Shenzhen. He has published over 70 articles in toptier academic journals and conferences. His current interests include pattern recognition, deep learning, biometrics, machine learning, and video analysis. He has served as a Co-Editor-in-Chief for the International Journal of Image and Graphics, an Associate Editor for the CAAI Transactions on Intelligence Technology, and an Editor for the Pattern Recognition and Artificial Intelligence.